AD-A065 626 MARYLAND UNIV COLLEGE PARK COMPUTER SCIENCE CENTER F/G 9/2
ANOTHER LOOK AT THE LONGLEY DATA SET.(U)
DEC 78 G W STEWART N00014-76-C-0391
UNCLASSIFIED TR-719 NL

1 OF 1
AD
A065626

code 432

LEVEL #1 NW

UNIVERSITY OF MARYLAND

COMPUTER SCIENCE CENTER

COLLEGE PARK, MARYLAND
20742

14 TR-719

9 Technical Rep[redacted]TR-719 11 Dec[redacted]78

6 Another Look at the Longley
Data Set.

12 31 p.

10 G. W. Stewart*

15 N00014-76-C-0391,
NSF-MSC76-03297

Abstract

This paper considers a linear regression problem involving economic data used by Longley [5] in a study of the performance of regression programs. The data set is notoriously difficult to handle computationally. In this paper, the singular value decomposition and the QR factorization are used to show that very small perturbations in the data render it colinear, thus accounting for the computational difficulties. Another analysis, based on coefficients that bound perturbations in the regression coefficients in terms of perturbations in the columns of the data, also shows the extreme sensitivity of the problem. An analysis is also given of a perturbation index, introduced by Beaton, Rubin, and Barone [1] to measure the sensitivity of regression problems. It is shown that the index is valid only for extremely large sample sizes and is not applicable to the Longley data set.

*This work was supported in part by the National Science Foundation and the Office of Naval Research under Contracts MSC76-03297 and N 00014-76-C-0391.

402 018 RLH

79 02 21 05

Another Look at the Longley Data Set

G. W. Stewart

1. Introduction

In a study of programs for solving regression problems, Longley [5] introduced a set of economic data on which several programs failed to compute acceptably accurate regression coefficients. Recently Beaton, Rubin, and Barone [1, hereafter referred to as BRB] showed that the regression coefficients are more affected by errors in the data itself than by rounding errors due to any reasonable computational scheme. In summary, they made the conservative assumption that the data was accurate in all reported figures and introduced pseudo-random perturbations uniformly distributed between -5 and 4.999 in the first unreported digit, so that the perturbed data rounded back to the original at the assumed number of significant digits. One thousand such data sets were generated, and regression coefficients were computed for each set, care being taken that rounding errors in the computation had negligible effects.

Each regression coefficient was found to vary in both sign and magnitude over the perturbed data sets. Moreover, the medians of the coefficients were not near the corresponding coefficients of the original problem, in spite of the fact that the perturbations in the data were symmetric. To explain this phenomena, a limiting solution, valid for large numbers of observations, was derived, along with a "perturbation index", which perportedly measures the sensitivity of the regression coefficients to errors in the data. Beaton, Rubin, and Barone conclude that "the use of stable algorithms and high precision is not likely to yield a valid answer without more accurate data" and that the perturbation index should "be used routinely to indicate the existance of severe instabilities in regression solutions."

The author agrees wholeheartedly with the first of these conclusions -- especially with the implication that what seem to be numerical problems may instead be symptoms of more fundamental statistical difficulties. However, there are easier ways to see that the Longley data set is a hard case than performing a large simulation experiment. One of the purposes of this paper is to present three ways, two of which provide a plausible explanation for the behavior of the medians of the regression coefficients. Specifically, we shall show that there are data sets with exact colinearities within the domain of perturbations considered by BRB. Moreover, we shall give reasons for believing that the perturbations introduced by BRB actually tend to make the problem better behaved, this bias accounting for the bias in the coefficients.

The third approach is to compute numbers that measure how sensitive the individual regression coefficients are to perturbations in the individual variables of the data set. These sensitivity coefficients immediately show that no accuracy can be expected in the regression coefficients in the presence of perturbations of the size considered by BRB.

The results of the sensitivity analysis are at variance with what the perturbation index implies about the coefficients. Accordingly, a section of this paper is devoted to an analysis of the asymptotic properties of the perturbation index, in which it is shown that it is a valid measure of sensitivity only when the number of observations is very large.

It will sometimes be convenient to cast the results of this paper in the language of norms. We shall use the vector 2-norm defined for any vector x by

$$\|x\| = (\Sigma x_i^2)^{1/2}.$$

For any matrix X we shall use either the Frobenius norm defined by

$$\|X\|_F = (\Sigma_{i,j} x_{ij}^2)^{1/2}$$

or the spectral norm defined by

$$\|X\|_2 = \sup_{\|x\|=1} \|X\,x\| .$$

The appearence of $\|X\|$ without a subscript in any statement means that the statement holds for either the Frobenius or the spectral norms. For a review of the properties of these norms see [6] .

I would like to thank Kathy Schmidt for her programming and computational help and David Hoaglin for his comments on a preliminary version of this paper.

2. The Longley data set

We consider the usual regression model

$$y = \beta_0 \underset{\sim}{1} + X\beta + e$$

1. Longley Data Set

	x_1	x_2	x_3	x_4	x_5	x_6	y
	GNP price deflator (x 10)	GNP	Unemploy-ment	Size of armed forces	Noninst. pop. 14 yrs. & over	Time	Total derived employment
Raw data	830	234,289	2,356	1,590	107,608	1947	60,323
	885	259,426	2,325	1,456	108,632	1948	61,122
	882	258,054	3,682	1,616	109,773	1949	60,171
	895	284,599	3,351	1,650	110,929	1950	61,187
	962	328,975	2,099	3,099	112,075	1951	63,221
	981	346,999	1,932	3,594	113,270	1952	63,639
	990	365,385	1,870	3,547	115,094	1953	64,989
	1000	363,112	3,578	3,350	116,219	1954	63,761
	1012	397,469	2,904	3,048	117,388	1955	66,019
	1046	419,180	2,822	2,857	118,734	1956	67,857
	1084	442,769	2,936	2,798	120,445	1957	68,169
	1108	444,546	4,681	2,637	121,950	1958	66,513
	1126	482,704	3,813	2,552	123,366	1959	68,655
	1142	502,601	3,931	2,514	125,368	1960	69,564
	1157	518,173	4,806	2,572	127,852	1961	69,331
	1169	554,984	4,007	2,827	130,081	1962	70,551
Beta	+15.0619	-0.0358	-2.0202	-1.0332	-0.0511	+1829.1515	

where $\underset{\sim}{1} = (1,1,\ldots,1)^T$ and X is a 16x6 matrix. The columns $x_1,\ldots,x_6$ of X contain observations of six independent variables, and the vector y contains observations of the dependent variable. Table 1 contains these observations*, along with the regression coefficients $\beta_0,\beta_1,\ldots,\beta_6$ (for further derived data, such as means, correlations, etc., see [1]).

We shall follow BRB in regarding this data as fixed and considering the effects of perturbations on the regression coefficients. Unless otherwise stated, the perturbations will be restricted to the interval $[-.5, .5]$ so that any perturbed data set rounds back to the original. This restriction is very conservative, since it is unlikely that any of the variables $x_1,x_2,\ldots,x_6$ are known to more than three figures.

We shall have occasion to work with the adjusted matrix X_a obtained from X by subtracting column means; i.e.

$$X_a = X - \underset{\sim}{1}m^T ,$$

where

$$m^T = \frac{\underset{\sim}{1}^T X}{16} .$$

Since the adjustment of X is by an additive factor, a perturbation in X corresponds to an identical perturbation in X_a. However, if we perturb X_a to get $\tilde{X}_a$ and form $\tilde{X} = \tilde{X}_a + \underset{\sim}{1}m^T$, the resulting $\tilde{X}$ adjusts back to $\tilde{X}_a$ if and only if

---

* For the variable x_1 we have reported the original data times ten, so that the perturbations defined below will have uniform ranges and variances.

$$1^T \tilde{X}_a = 0. \tag{2.1}$$

Thus if we wish to induce perturbations in X by perturbing X_a we must take care that (2.1) is satisfied. This point will prove important in the next two sections.

3. Singular value analysis

It is well known (e.g. see [6]) that for any $n \times p$ matrix X with $n \geq p$, there are orthogonal matrices U and V such that

$$U^T X V = \begin{pmatrix} M \\ 0 \end{pmatrix}, \tag{3.1}$$

where

$$M = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_p)$$

and

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_p \geq 0.$$

The decomposition (3.1) is called the <u>singular value decomposition</u> of X. The numbers $\mu_1, \mu_2, \ldots, \mu_p$ are the singular values of X and the columns of U and V are respectively the left and right singular vectors of X. If U is partitioned in the form

$$U = (U_1, U_2)$$

where U_1 is $n \times p$ then

$$X = U_1 M V^T, \tag{3.2}$$

an expression which is sometimes called the singular value factorization of X.

The singular value decomposition has an important approximation property. Given the integer $k \le p$, let

$$\tilde{M} = \mathrm{diag}(\mu_1,\ldots,\mu_k,0,\ldots,0),$$

and let

$$\tilde{X} = U_1 \tilde{M} V^T. \tag{3.3}$$

Then $\tilde{X}$ has rank less than or equal to k, and for any $n \times p$ matrix Y of rank less than or equal to k

$$\|X - \tilde{X}\|^2 \le \|X - Y\|^2.$$

Thus $\tilde{X}$ is a matrix of rank less than or equal to k that is nearest to X in the least squares sense, and $\tilde{X}$ is easily computable from the singular value factorization of X.

This method of obtaining nearby matrices with colinearities may be applied to the Longley data set. If we compute the singular value decomposition of the matrix X_a for the Longley data set (the LINPACK code SSVDC was used [3]), we get the following sequence of singular values, rounded to two places:

$$3.9\cdot 10^5,\ 4.7\cdot 10^3,\ 1.7\cdot 10^3,\ 1.3\cdot 10^3,\ 3.7\cdot 10^1,\ 6.7\cdot 10^{-1}. \tag{3.4}$$

The smallest singular value is near the error range described in §2. Accordingly, we set it to zero and compute $\tilde{X}_a$ in analogy with (3.3) as

$$(3.5) \qquad \tilde{X}_a = U_1\tilde{M}V^T$$

and then form

$$\tilde{X} = \tilde{X}_a + \underset{\sim}{1}m^T.$$

In order for this process to be legitimate, the condition (2.1) must be satisfied, so that the adjusted $\tilde{X}$ is the rank deffficient matrix $\tilde{X}_a$. Since M is nonsingular, it follows from (3.2), with X replaced by X_a that

$$U_1 = X_a V M^{-1}$$

Since $\underset{\sim}{1}^T X_a = 0$, it follows that $\underset{\sim}{1}^T U_1 = 0$ and

$$\underset{\sim}{1}^T \tilde{X}_a = \underset{\sim}{1}^T U_1 \tilde{M} V^T = 0,$$

which is just the condition (2.1).

The first and sixth columns of X are reproduced to eight figures in Table 2 (the deviations of the other columns were below the level of rounding error). The largest deviation from X occurs in the year 1951 and has a value of 0.4196. Thus the perturbations are well within the range described in Section 2. It follows that, for all one knows, the "true" values of the Longley data set could harbor an exact colinearity. In particular, within the domain of matrices treated by BRB, there are points where the regression coefficients fail to exist, and near these points the coefficients can become arbitrarily large. Under the circumstances, it is not surprising that the coefficients behave erratically.

However, we believe that the shifting of the centers of the coefficients observed by BRB is due to the surprising fact that the perturbations tend to move the problem <u>away</u> from the singularities just mentioned.

2. Rank Deficient Approximations to X

SVD x_1	SVD x_6	QR x_6
829.99998	1946.9943	1946.9942
885.00009	1948.0257	1948.0257
881.99992	1948.9759	1948.9759
894.99996	1949.9891	1949.9891
962.00146	1951.4196	1951.4196
981.00058	1952.1662	1952.1662
989.99931	1952.8004	1952.8004
999.99956	1953.8733	1953.8733
1011.9999	1954.9580	1954.9580
1045.9991	1955.7386	1955.7386
1083.9991	1956.7529	1956.7529
1107.9999	1957.9848	1957.9848
1126.0004	1959.1098	1959.1098
1141.9998	1959.9358	1959.9358
1157.0004	1961.1113	1961.1113
1169.0006	1962.1646	1962.1646

To see how this may happen, we must look at the effects of perturbations in a matrix X on its smallest singular value. Let $\tilde{X} = X + E$, where we assume that the elements of E are uncorrelated with mean zero and common variance σ^2. From (3.2) it is easy to see that the eigenvalues of X^TX are $\mu_1^2, \mu_2^2, \ldots, \mu_p^2$ with corresponding eigenvectors $v_1, v_2, \ldots, v_p$, where v_j is the j-th column of V. It follows that the square $\tilde{\mu}_p^2$ of the smallest singular value of $\tilde{X}$ is the smallest eigenvalue of

$$(X + E)^T(X + E) = X^TX + X^TE + E^TX + E^TE.$$

The first order approximation to $\tilde{\mu}_p^2$ is given by $v_p^T(X + E)^T(X + E)v_p$ (e.g. see [6]). If we use the facts that $Xv_p = \mu_p u_p$ and $X^Tu_p = \mu_p v_p$, then

$$\begin{aligned}
\tilde{\mu}_p^2 &\doteq v_p^TX^TXv_p + v_p^TX^TEv_p + v_p^TE^TXv_p + v_p^TE^TUU^TEv_p \\
&= \mu_p^2 + 2\mu_p u_p^TEv_p + \sum_{i=1}^{n}(u_i^TEv_p)(u_i^TEv_p) \\
&= \mu_p^2 + 2\mu_p u_p^TEv_p + (u_p^TEv_p)^2 + \sum_{\substack{i=1 \\ i\neq p}}^{n}(u_i^TEv_p)^2 \qquad (3.6) \\
&= (\mu_p + u_p^TEv_p)^2 + \tau^2,
\end{aligned}$$

where

$$\tau^2 = \Sigma_{\substack{i=1 \\ i\neq p}}^{n}(u_i^TEv_p)^2.$$

From the distributional assumptions on E and the orthonormality of the u_i and the v_j, it follows that

(3.7) $$E(\tau^2) = (n-1)\sigma^2.$$

Thus (3.6) partitions the first order approximation to $\tilde{\mu}_p^2$ into two terms, one the square of a term deviating from μ_p by a quantity with standard deviation σ and the other a sum of squares with mean $(n-1)\sigma^2$. As long as μ_p is sufficiently larger than σ, the fluctuations in $\tilde{\mu}_p$ are almost entirely due to the first term. But as μ_p approaches σ, the second term will dominate, and tend to increase the value of μ_p. To summarize this informal argument: <u>if the elements of a matrix X are perturbed by quantities nearly equal to its smallest singular value, the perturbations will tend to increase that singular value</u>.

In the case of the BRB experiments, we have $\sigma^2 = 1/12$ and $\mu_6^2 \cong .45$. From (3.7) it follows that

$$E(\tau^2) = 1.25.$$

Thus the τ^2 term dominates, and the effect of the perturbations is for the most part to produce a better behaved problem with μ_6 increased. We believe that this bias toward nicer problems is the cause of the bias in the perturbed coefficients observed by BRB.

In the foregoing we have taken care to scale the columns of X so that the presumed uncertainties in the data are all equal. This has the effect of making the singular values readily interpretable in terms of the matrix X; the suppression of a singular value less than the uncertaintity will cause the elements of X to be perturbed by quantities of the same magnitude. On the other hand, if the variables were scaled so that the uncertainties were disparate, the suppression of a small singular value could overwhelm a still smaller uncertainty in a particular column. We mention this point because it is a common practice to scale X_a so that its columns have norm one (in which case the columns of V are the principal components of the problem). Whatever the merits of this approach in other circumstances, it is clearly not the thing to do here.

4. Analysis via the QR decomposition

A comparison of Table 2 with Table 1 shows that the perturbations introduced by the singular value analysis occur mostly in the time variable x_6. This suggests the possibility of obtaining a singular perturbation of X by changing only the sixth column. In this section we shall show how the QR decomposition may be used to find such a perturbation.

Given any $n \times p$ matrix X with $n \geq p$, there is an orthogonal matrix Q such that

$$Q^T X = \begin{pmatrix} R \\ 0 \end{pmatrix} \tag{4.1}$$

where R is upper triangular (e.g. see [6]). This decomposition of X is called the QR decomposition. If we write $Q = (Q_1, Q_2)$, where Q_1 is $n \times p$, then it follows from (4.1) that

$$X = Q_1 R, \tag{4.2}$$

and (4.2) is called the QR factorization of X.

The QR decomposition is a useful computational and theoretical tool in linear regression; however, for our purposes we need only the following approximation theorem, which appears to be new.

<u>Theorem</u> 3.1. In the QR decomposition (4.1), suppose that R is nonsingular. Let $\tilde{R}$ be obtained from R by setting $r_{pp} = 0$, and let

$$\tilde{X} = Q_1\tilde{R}. \tag{4.3}$$

Then $\tilde{X}$ differs from X only in its p-th column, and $\text{rank}(\tilde{X}) = p-1$. Moreover, if Y is an $n \times p$ matrix that differs from X only in its p-th column and satisfies $\text{rank}(Y) \leq p-1$, then

$$\|X - \tilde{X}\| \leq \|X - Y\| . \tag{4.4}$$

<u>Proof</u>. By construction R and R differ only in their (p,p)-elements. Hence, X and $\tilde{X}$ differ only in their p-th columns. Moreover, $\tilde{R}$ is of rank p-1, and therefore so is $\tilde{X}$.

To establish (4.4), let R be partitioned in the form

$$R = \begin{pmatrix} R' & r \\ 0 & r_{pp} \end{pmatrix}$$

where R' is of order p-1. Let y_p denote the p-th column of Y, and partition $z = Q^T y_p$ in the form

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

where z_1 is a (p-1)-vector and z_2 is a scalar. Then

$$Q^T\tilde{X} = \begin{pmatrix} R' & r \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$Q^T Y = \begin{pmatrix} R' & z_1 \\ 0 & z_2 \\ 0 & z_3 \end{pmatrix}$$

It follows that

$$\|X - \tilde{X}\|^2 = r_{pp}^2 . \tag{4.5}$$

Now for Y to have rank p-1, the quantities z_2 and z_3 must be zero. Hence

$$\|X - Y\|^2 = \|r - z_1\|^2 + r_{pp}^2 . \tag{4.6}$$

The inequality (4.4) follows from (4.5) and (4.6).

The application of this theorem to the Longley data set is similar to the singular value analysis. The QR decomposition of the matrix X_a was computed by the LINPACK routine SQRDC [3]. The element

$$r_{66} = 0.6693051$$

of R was set to zero to give $\tilde{R}$ and $\tilde{X}_a$ computed in analogy with (4.3) as

$$\tilde{X}_a = Q_1\tilde{R} .$$

An argument similar to the one in the last section establishes that $\underset{\sim}{1}^T Q_1 = 0$. Hence $\tilde{X}_a$ satisfies (2.1), and we may add means as usual to get $\tilde{X}$. The sixth column of $\tilde{X}$, which is the only one that has been altered, has been appended to Table 2. The largest deviate corresponds to the year 1951 and has a value of 1951.420, so that $\tilde{X}$ again lies within the range of perturbations considered by BRB.

We observed in connection with the singular value decomposition that perturbations could tend to make a problem better behaved. Much the same thing can occur with errors introduced into a single column. Specifically, let X have the QR decomposition (4.1) and let $\tilde{X}$ be obtained from X by adding to x_p a vector e whose elements are uncorrelated with mean zero and common variance σ^2. Let $f = Q^T e$. Then if we partition f in the form

$$f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} ,$$

where f_1 is a (p-1)-vector and f_2 is a scalar, we have in the notation used above,

$$Q^T \tilde{X} = \begin{pmatrix} R' & r + f_1 \\ 0 & r_{pp} + f_2 \\ 0 & f_3 \end{pmatrix} .$$

It follows that the (p,p)-element of $\tilde{R}$ satisfies

$$\tilde{r}_{pp}^2 = (r_{pp} + f_2)^2 + \|f_3\|^2 ,$$

and

$$E(\|f_3\|^2) = (n-p)\sigma^2 .$$

If r_{pp} is near σ in magnitude, the term $\|f_3\|^2$ will tend to dominate and increase r_{pp}.

For perturbations in the variable x_6 of the Longley data set we have

$$(n-p)\sigma^2 = \frac{10}{12} = .83$$

which clearly dominates $r_{pp}^2 = .45$. Although this analysis does not strictly apply to the perturbations considered by BRB, since it assumes the other variables are not perturbed, it none the less gives a fair indication of what is going on. In ten simulations, done by the author for other purposes, it was observed that the average value of r_{pp}^2 was 1.2, which is in fair agreement with expectation 1.37 of $\tilde{r}_{pp}^2$ in (4.7).

For this data set, the QR decomposition yields much the same results as the singular value decomposition. However, this is in part due to the fortuitous ordering of the variable x_6; another ordering of the columns could give different results. In general, it may be necessary to inspect r_{pp} for different orderings. There is no need to examine all 2^{p-1} orderings, since the value of r_{pp} depends only on the variable that is placed last and not on the ordering of the other variables. Efficient algorithms exist to determine these p different values of r_{pp} after R has been computed once for a specific ordering [3, Ch.10].

5. Sensitivity coefficients

The results of the last two sections suggest that the regression coefficients for the Longley data set will be extremely sensitive to perturbations in the variable x_6 and, to a lesser extent, in the variable x_1. For sufficiently small perturbations, we can make this precise by computing linear approximations to the perturbations in the regression coefficients. In this section we shall summarize the results of such an approach. The reader will find details in [4] or [8].

In a general regression model with regression matrix X, assume that S is of full rank so that the vector of least squares coefficients is given by

(5.1) $$\beta = (X^T X)^{-1} X^T y = CX^T y \equiv X^{\dagger} y ,$$

where for later use we have set

$$C = (X^T X)^{-1}$$

and

$$X^{\dagger} = (X^T X)^{-1} X^T$$

($X^{\dagger}$ is the pseudo-inverse of X). From (5.1) it is evident that if $\tilde{X}$ is restricted to a sufficiently small neighborhood of X, then $\tilde{\beta} = \tilde{X}^{\dagger} y$ is a differentiable function of $\tilde{X}$. In particular, if we write $\tilde{\beta}_i$ as a function of the j-th column of $\tilde{X}$, say

$$\tilde{\beta}_i = f_{ij}(\tilde{x}_j) ,$$

then $\tilde{\beta}_i$ can be expressed in the form

$$\tilde{\beta}_i = \beta_i + f'_{ij}(x_j)(\tilde{x}_j - x_j) + O(\|\tilde{x}_j - x_j\|^2) , \tag{5.2}$$

where the row vector $f'_{ij}(x_j)$ is the gradient of f_{ij} evaluated at x_j. It turns out that there is an easy expression for $f'_{ij}(x_j)$ and its norm γ_{ij}.

<u>Theorem</u> 4.1 4,8 . Let $x_i^{(+)}$ denote the i-th row of X^+, and let

$$r = y - X\beta .$$

Then

$$f'_{ij}(x_j) = -\beta_j x_i^{(+)} + c_{ij} r^T$$

and

$$\gamma_{ij}^2 \equiv \| f'_{ij}(x_j) \|^2 = \beta_j^2 c_{ii} + \|r\|^2 c_{ij}^2 .$$

There are two ways in which this theorem can be applied. In the first place, it follows from (5.2) that

$$|\tilde{\beta}_i - \beta_i| \le \gamma_{ij} \|\tilde{x}_j - x_j\| + O(\|\tilde{x}_j - x_j\|^2) .$$

Thus if we can place a bound on the size of the perturbation $\tilde{x}_j - x_j$ in x_j and the perturbation is sufficiently small*, then $\gamma_{ij} \| x_j - x_j \|$

---

*This will be true if $\| X^+ \| \| \tilde{x}_j - x_j \|$ is significantly less than one, say less than 0.2, a result which can be derived from theorems in [7] .

estimates the perturbation in β_i due to the perturbation in x_j. For this reason we shall call γ_{ij} a <u>sensitivity</u> <u>coefficient</u>.

A second approach is to make distributional assumptions about the components of $\tilde{x}_j - x_j$, say that they are independently distributed with means zero and common variances σ^2. Then the variance of the approximation

$$\tilde{\beta}_i' = \beta_i + f_{ij}'(x_j)(\tilde{x}_j - x_j) \tag{5.2}$$

is $\gamma_{ij}^2\sigma^2$, so that again γ_{ij} estimates the variability of $\tilde{\beta}_i$ due to perturbations in x_j. However, some care is required here. If the distribution of $\tilde{x}_j - x_j$ is continuous and nonzero at a singularity of $\tilde{X}$, then we cannot guarantee that the moments of $\tilde{\beta}_j$ exist. This will always be the case if $\tilde{x}_j - x_j$ is normally distributed. Intuition suggests that if σ^2 is small enough then $\tilde{\beta}_i'$ will accurately approximate $\tilde{\beta}_i$ except in a region of low probability, so that $\gamma_{ij}^2\sigma^2$ will adequately describe the variability of $\tilde{\beta}_i$; however, this area needs further study.

The sensitivity coefficients can easily be computed from quantities normally generated in the course of solving regression problems. We have done this for the matrix X_a obtained from the Longley data set and the adjusted vector $y_a = y - (\underset{\sim}{1}^T y)\underset{\sim}{1}/16$. Since the regression coefficients differ widely in magnitude, we report γ_{ij}/β_i in Table 3. These scaled coefficients measure the sensitivity of the relative error $|\tilde{\beta}_i - \beta_i|/|\beta_i|$; if this error if less than 10^{-s} then $\tilde{\beta}_i$ and β_i agree to about s significant figures.

3. Relative sensitive coefficients γ_{ij}/β_i

i \ j	1	2	3	4	5	6
1	$4.7\cdot10^{-1}$	$1.4\cdot10^{-3}$	$4.0\cdot10^{-2}$	$2.0\cdot10^{-2}$	$8.3\cdot10^{-3}$	$3.4\cdot10^{1}$
2	$5.1\cdot10^{-2}$	$3.3\cdot10^{-4}$	$7.5\cdot10^{-3}$	$3.3\cdot10^{-3}$	$1.7\cdot10^{-3}$	$6.5\cdot10^{-3}$
3	$1.1\cdot10^{-2}$	$8.1\cdot10^{-5}$	$2.0\cdot10^{-3}$	$8.8\cdot10^{-4}$	$4.1\cdot10^{-4}$	$1.7\cdot10^{0}$
4	$6.1\cdot10^{-3}$	$4.0\cdot10^{-3}$	$1.5\cdot10^{-3}$	$8.3\cdot10^{-4}$	$9.4\cdot10^{-5}$	$1.3\cdot10^{0}$
5	$2.4\cdot10^{-1}$	$1.3\cdot10^{-3}$	$3.3\cdot10^{-2}$	$1.5\cdot10^{-2}$	$9.9\cdot10^{-3}$	$2.7\cdot10^{1}$
6	$4.1\cdot10^{-3}$	$7.2\cdot10^{-5}$	$1.9\cdot10^{-3}$	$8.9\cdot10^{-4}$	$2.1\cdot10^{-4}$	$1.8\cdot10^{0}$

The coefficients confirm our conclusions about the sensitivity of the problem to perturbations in x_1 and x_6. For example, the coefficient γ_{16}/β_1 is 34. If we consider perturbations introduced by rounding the elements of x_6 in the s-th place beyond the decimal, then the maximum such perturbation is $\pm 5\cdot 10^{-s}$. If follows that $\|\tilde{x}_6 - x_6\| \leq 20\cdot 10^{-s}$; hence

$$\frac{|\tilde{\beta}_1 - \beta_1|}{|\beta_1|} \leq 34\cdot 20\cdot 10^{-s}$$

To be sure of one figure of accuracy in $\tilde{\beta}_1$, we must have $34\cdot 20\cdot 10^{-s} \lesssim .1$ or $s \geq 4$. The corresponding perturbation of $\pm 5\cdot 10^{-4}$ years amounts to about ± 4.4 hours. Although this is a worst case analysis, it reflects the extreme sensitivity of β_1 to perturbations in x_6; a probabistic analysis would give only slightly less dramatic results. The sensitivity coefficients also show that β_1 and β_5 are quite sensitive to perturbations in x_1.

We must insert a word of caution here. The results of the last three sections all agree in condemning the variable x_6 as a trouble maker, and to a lesser extent the variable x_1. It is tempting to conclude that all will be well if we exclude x_6 and x_1 from the model. However, the sensitivity of the coefficients to x_1 and x_6 is a function of the entire model. There is no reason to expect that either x_6 or x_1 cannot behave themselves in a reduced model. The techniques we have described in this paper are designed to detect trouble, not to remedy it, and we discourage their naive application to the variable selection problem.

6. Limitations of a perturbation index.

The first order perturbation theory of the last section is sharp in proportion as the variance is small. A different approach would fix the variance of the errors and investigate what happens as n becomes large. This case has been analyzed in [1] and [2]. In this section we shall be concerned with how large n must be for the analyses to be applicable.

The basic results are derived as follows. We begin with a sequence of regression problems with full rank $n \times p$ matrices X_n $(n = 1,2,\ldots)$ and observation vectors y_n $(n = 1,2,\ldots p)$. The coefficient vectors b_n are given by

$$b_n = (X_n^T X_n)^{-1} X_n^T y_n \ .$$

We suppose further that there is a limit problem in the sense that there is a positive definite matrix A, a p-vector c, and a scalar η^2 such that

$$\begin{aligned} &\lim_{n\to\infty} n^{-1} X_n^T X_n = A \ , \\ (6.1)\qquad &\\ &\lim_{n\to\infty} n^{-1} X_n^T y_n = c \ , \end{aligned}$$

and

$$(6.2)\qquad \lim_{n\to\infty} n^{-1}\, y_n{}^2 = \eta^2 .$$

It follows that

$$\lim_{n\to\infty} b_n = A^{-1} c \equiv b.$$

Now suppose that we are actually given the matrices

$$\tilde{X}_n = X_n + E_n,$$

where the elements of E_n are assumed to be uncorrelated with mean zero and common variance σ^2. The coefficient vector obtained by working with $\tilde{X}_n$ instead of X_n is given by

$$(6.3)\qquad \begin{aligned} b_n &= (\tilde{X}_n^T\tilde{X}_n)^{-1}\tilde{X}_n^T y \\ &= \left[n^{-1}(X_n^T X_n + X_n^T E_n + E_n^T X_n + E_n^T E_n)\right]^{-1} n^{-1}(X_n^T y_n + E_n^T y_n) \end{aligned}$$

The limits in probability of the terms in the right hand side of (6.3) can easily be evaluated. From the assumptions on E_n we have immediately that

$$\operatorname*{plim}_{n\to\infty} n^{-1}(E_n^T E_n) = \sigma^2 I.$$

Next, from (6.1) it follows that if $x_i^{(n)}$ denotes the i-th column of X_n, then

$$(6.4)\qquad \lim_{n\to\infty} \frac{\|x_i^{(n)}\|^2}{n} = a_{ii}.$$

Hence $n^{-1/2}X_n$ is bounded and

$$\operatorname*{plim}_{n\to\infty} \frac{X_n^T E_n}{n} = 0.$$

Finally, from (6.2) it follows that $n^{-1/2}y_n$ is bounded and

$$\operatorname*{plim}_{n\to\infty} \frac{E_n^T y_n}{n} = 0.$$

Hence

$$\operatorname*{plim}_{n\to\infty} \tilde{b}_n = (A + \sigma^2 I)^{-1} c.$$

Equation (6.5) shows clearly that $\operatorname{plim} \tilde{b}_n$ differs from the true solution b by quantities that depend on the variance of E. We may obtain specific bounds for this difference by applying results from standard matrix perturbation theory (e.g. see [6]). Specifically, if

$$\sigma^2 \| A^{-1} \| < 1 \tag{6.6}$$

then $(A + \sigma^2 I)$ is nonsingular and

$$\frac{\| b - \operatorname{plim} \tilde{b}_n \|}{\| b \|} \le \frac{\sigma^2 \| A^{-1} \|}{1 - \sigma^2 \| A^{-1} \|} . \tag{6.7}$$

Since $\operatorname{trace}(A^{-1}) \ge \| A^{-1} \|$, we may replace the condition (6.6) by

$$\sigma^2 \operatorname{trace}(A) < 1$$

and the bound (6.7) by

$$\frac{\| b - \operatorname{plim} \tilde{b}_n \|}{\| b \|} \le \frac{\sigma^2 \operatorname{trace}(A^{-1})}{1 - \sigma^2 \operatorname{trace}(A^{-1})} . \tag{6.8}$$

The right hand side of (6.7) or (6.8) is a relative error in the vector β. If it is of order 10^{-s}, then the largest components of plim $\tilde{b}_n$ will be in error in about their s-th digit. The bounds may then be interpretated as saying that if either $\sigma^2 \| A^{-1} \|$ or $\sigma^2 \mathrm{trace}(A^{-1})$ is near one, the plim of $\tilde{b}_n$ may differ entirely from b. For this reason, BRB call $\sigma^2 \mathrm{trace}(A^{-1})$ the perturbation index for the problem and recommend that it be monitored to determine the sensitivity of the problem to errors in the variables.

To compute the perturbation index for the Longley data set, we approximate $A \cong X_a^T X_a/16$. Now

$$\mathrm{trace}(X_a^T X_a)^{-1} = \frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \ldots + \frac{1}{\mu_6^2},$$

where the μ_i are the singular values displayed in (3.4). Thus

$$\mathrm{trace}(A^{-1}) \cong 36.8$$

For uniform errors in the first unreported figure, we have $\sigma^2 = 1/12$, so that the perturbation index is about three, which gives ample warning of trouble.

However, if we consider errors in the second unreported figure, we have $\sigma^2 = 1/1200$, and hence the perturbation index is about 0.03, a value which promises reasonable accuracy in plim $\tilde{\beta}_n$. On the other hand the sensitivity coefficients suggest that the relative error in the coefficient β_1 due to perturbations of this kind in x_6 will be of order of magnitude

$$\frac{\gamma_{16}}{\beta_1}\sigma \cong 34.\frac{1}{\sqrt{1200}} \cong .98$$

Thus we can expect no accuracy in β_1, in spite of the small perturbation index.

The cause of the difficulty is that n must be very large for $\hat{b}_n$ to approximate its plim with any degree of certainty. Returning to (6.3), we see that replacing the matrix $n^{-1}(X_n^T E_n + E_n^T X_n)$ by its plim of zero can only be justified if it is small in probability compared with the plim of $n^{-1}E_n^T E_n$. In particular, the variance of a diagonal element of $n^{-1}(X_n^T E_n + E_n^T X_n)$ is

$$\frac{4}{n^2} E\left[(x_i^{(n)T} e_i^{(n)})^2\right] = \frac{4\sigma^2}{n} \frac{\|x_i^{(n)}\|^2}{n} \cong \frac{4\sigma^2}{n} a_{ii}.$$

This variance must be small compared with the square of the corresponding diagonal element of plim $n^{-1}E_n^T E_n$, which is σ^4. Hence n must at least satisfy

$$\frac{4\sigma^2 a_{ii}}{n} < \sigma^4$$

or

$$n > 4\frac{a_{ii}}{\sigma^2}. \tag{6.9}$$

The number $\sigma/\sqrt{a_{ii}}$ is a measure of the relative size of the perturbations in the i-th column (if the data has been adjusted, $\sqrt{a_{ii}}$ is approximately the standard deviation of the elements of the i-th column). For example, if the data is accurate to three figures, then $\sigma/\sqrt{a_{ii}} \cong 10^{-3}$ and from (6.9) it follows that n must be at least four million before the analysis leading to the perturbation index is to be trusted. If we are concerned with rounding errors on, say, a computer carrying eight decimal digits, then $\sigma/\sqrt{a_{ii}} \cong 10^{-8}$ and n must be in the quadrillions.

As far as the Longley data is concerned, the largest standard deviation occurs for the variable x_2 and is about 10^5. Taking $\sigma^2=1/12$, we must have

$$n > 4\cdot 12\cdot 10^{10} = 4.8\cdot 10^{11},$$

a criterion which the sixteen observations in the Longley data set fall short of satisfying.

7. Conclusions.

Although we have confined our attention to the Longley data set in this paper, the techniques that we have used are quite general. If one can estimate the sizes of the errors in the variables, then the singular value decomposition provides a way of seeing if they can have disastrous effects (we again stress the need for proper scaling of X). The QR decomposition allows one to search for particularly offensive columns.

Perhaps most useful of all are the sensitivity coefficients. Being derived from a linearization of the problem, they are not valid for large errors; however, if a problem is locally sensitive, then large errors are unlikely to correct the difficulty. We add that efficient software for implementing these techniques exists, and that, properly done, none of them will cause an order of magnitude change in the costs of computation.

As regards the perturbation index, we recommend that its use be eschewed. Although a perturbation index near to or greater than one is certainly a sign of trouble, it can be misleadingly small. Moreover, it measures effects that, in most practical circumstances, can be seen only when the sample size is astronomically large.

References

1. A. E. Beaton, D. B. Rubin, and J. L. Barone, The acceptability of regression solutions: another look at computational stability, J. Amer. Stat. Assoc. 71 (1976) 158-168.

2. R. B. Davies and B. Hutton, The effect of errors in the independent variables in linear regression, Biometrika 62 (1975) 383-391.

3. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, The LINPACK Users Guide, SIAM, Philadelphia (1979).

4. S. D. Hodges and P. G. Moore, Data uncertainties and least squares regression, Appl. Stat. 21 (1972) 185-195.

5. J. W. Longley, An appraisal of least squares programs for the electronic computer from the point of view of the user, J. Amer. Stat. Assoc. 62 (1967) 819-841.

6. G. W. Stewart, Introduction to Matrix Computations, Academic Press, New York (1974).

7. _______, On the perturbation of pseudo-inverses, projections, and linear least squares problems, SIAM Rev. 19 (1977) 634-662.

8. _______, Sensitivity coefficients for the effects of errors in the independent variables in a linear regression, University of Maryland, Computer Science Technical Report TR-571 (1977)